Analyzing Graduate Admission Data

Luis J. Ramirez
Dept. of Computer Science
Tecnologico de Monterrey
Monterrey, Mexico
A01280418@itesm.mx

Miguel A. Cruz-Gomez
Dept. of Computer Science
Tecnologico de Monterrey
Monterrey, Mexico
A01037093@itesm.mx

***Abstract*—We will use the Graduate Admissions dataset to analyze which are the most important things to consider for students applying to graduate school.**

***Index Terms*—admissions, data science, graduate school**

I. Introduction

Graduate school admissions can be very competitive, so knowing which characteristics do school consider the most for accepting students can be a good advantage for the people applying. It can also be very useful to the university as it could help automate and speed the admission process. We know that schools consider things such as exams, GPA, relevant experience, recommendation letters, etc, but we do not know how much of this influences on the final decision. We will analyze the Graduate Admissions dataset.

This project will be developed using AutoML platform by H2O, which is an open-source platform that helps automate the generation of machine learning models.

II. Methods and Data

A. Data

For this work, we consumed the Graduate Admissions dataset provided by Acharya in Kaggle. (https://www.kaggle.com/datasets/mohansacharya/graduate-admissions?resource=download). [1]

B. Exploratory Data Analysis

We found this dataset useful to build a proxy to the current situation on graduate student admission process. We defined the target variable "Chance of admit"; as it is shown, there are no missing data so it is not necessary to impute values; also, it is noted that some variables are int64 while others are float64 type, which are suitable data types for the models selected in the further analysis. As a proxy to the predictive power, we selected the Pearson's correlation to select the variables that are more likely to explain the dataset variance. Given the correlation matrix to model these relationships and comparing with the target variable, the feature with the highest absolute Pearon correlation is CGPA, followed by GRE Score, TOEFL Score and University Rating. Other categorical variables like Research have very poor correlation compared to the other features so they were excluded in the selection process. The selected features to build this model were: "CGPA", "GRE Score" , "University Rating", "SOP" and "LOR"; we excluded some variables like "TOEFL Score" as it is strongly correlated to the selected variables and this might have an impact on the model's performance.

Furthermore, we created representative features given our personal criteria; the feature "GRE Score / University Rating" as both are two of the most correlated variables with the feature; also we built the categorical variable "High GRE Score" that indicates if the corresponding "GRE Score" is located at Q1 of the GRE Score distribution; adding these variables increases the probability of convergence for the optimization process in a supervised learning approach.

C. Methods

We will use the tool AutoML from H2O to automate the generation of models and hyperparameter tuning in order to find the best model. The algorithms that will be tested are Stacked Ensembles (SE), Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), Generalized Linear Model (GLM), and Extreme Gradient Boosting (XGB).

Different metrics are evaluated, such as Mean Squared Error (MSE), Root-Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Mean Squared Log Error (RMSLE), and Mean Residual Deviance (MRD), we will be using MRD to select the best model.

5-fold cross validation was done to evaluate the performance of the different models.

III. Results

We created a total of 58 models by doing hyperparameter tuning in the 5 algorithms mentioned above. This models were tuned using AutoML. We evaluated all the metrics for all of them and ranked them in respect to MRD. We also calculated the average performance for each algorithm, to get a sense of how the algorithms performed on average independently of the hyperparameter tuning. We can see those results in Figures 2, 3 and 4. Figure 5 shows the features importances, as it another quantitative measure of the feature influence over the variability of the target variable. These last results are congruent with our exploratory data analysis conclusions, as CGPA and GRE Score happened to be strong predictors of the probability of being admit.

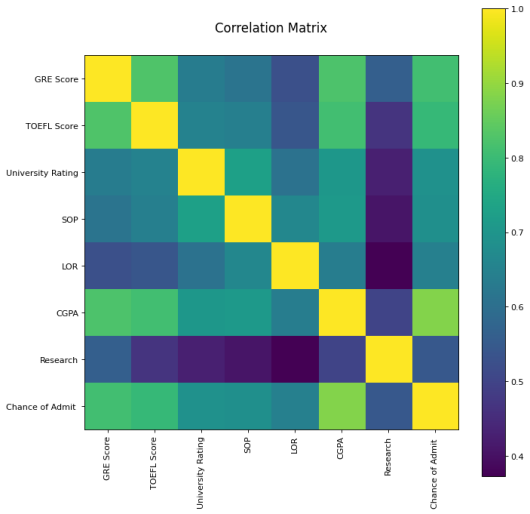

Fig. 1. Correlation matrix of variables

model_id	MRD	RMSE	MSE	MAE	RMSLE	Algorithm
GBM_model_11	0.003	0.051	0.003	0.038	0.032	GBM
GBM_model_27	0.003	0.052	0.003	0.038	0.032	GBM
StackedEnsemble_BestOfFamily_8	0.003	0.052	0.003	0.038	0.032	StackedEnsemble
StackedEnsemble_BestOfFamily_4	0.003	0.052	0.003	0.038	0.032	StackedEnsemble
GBM_model_28	0.003	0.052	0.003	0.038	0.032	GBM
StackedEnsemble_BestOfFamily_3	0.003	0.052	0.003	0.038	0.032	StackedEnsemble
StackedEnsemble_AllModels_3	0.003	0.052	0.003	0.038	0.032	StackedEnsemble
StackedEnsemble_BestOfFamily_7	0.003	0.052	0.003	0.037	0.032	StackedEnsemble
GBM_model_14	0.003	0.052	0.003	0.038	0.032	GBM
GBM_model_24	0.003	0.052	0.003	0.039	0.032	GBM

Fig. 2. Results of the top 10 algorithms

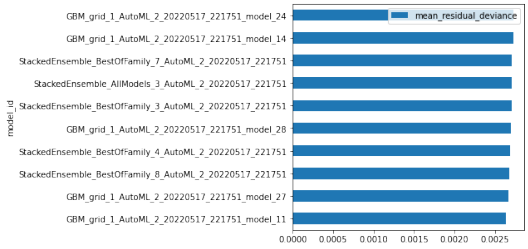

Fig. 3. MRD of top 10 algorithms

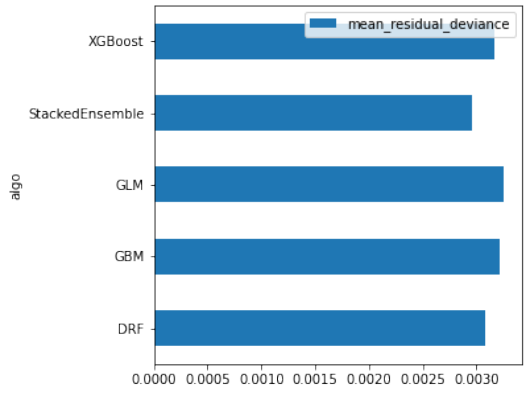

Fig. 4. Average MRD by algorithm

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
0	CGPA	13.390168	1.000000	0.362612
1	GRE Score	8.400928	0.627395	0.227501
2	TOEFL Score	5.562245	0.415398	0.150628
3	University Rating	3.526217	0.263344	0.095491
4	SOP	2.513360	0.187702	0.068063
5	C1	2.175205	0.162448	0.058906
6	LOR	0.774226	0.057821	0.020966
7	Research	0.584680	0.043665	0.015833

Fig. 5. Feature importances

IV. Discussion

In Figure 2, a table shows the results of the predictive analysis. Several metrics were used to describe the performance of the models and the results are grouped by algorithm. In Figure 3, The mean residual deviance is shown for the top 10 performing models. In Figure 4, the same metric is bar plotted by algorithm, to illustrate which algorithm suited better the data set.

In figure 6, we can see the results obtained originally on this dataset. Comparing our results, we can see that ours are better, as literature shows the best MSE to be at 0.005, vs our best result of 0.003. This mean we reduced the MSE in around 40%. [1]

Regression Models	MSE	R2 Score
Linear Regression	0.00480149	0.72486310
Support Vector Regression	0.00724206	0.64401301
Decision Tree Regression	0.00874299	0.50134421

Fig. 6. Results from Acharya 2019.

V. Conclusion

The results shown in Figure 2 illustrate that the Gradient Boosting Model outperforms the overall set of models, with a RMSE of 0.05; relative to the scale of the target variable, it represents a percent error of 5%. The features selected in the previous stage of the project happened to be strong predictors of the chance of admission.

We observe that GBM and SE outperform all of the other models we tried (DRF, XGB, GLM), and also outperform the other models that have been tested in the literature (Linear Regression, SVM and Decision Tree Regression).

This work also helps us see that open-source tools such as the AutoML platform from H2O can be very useful in the development of ML models, as it helps you train better models in less time and improve performance.

As a general conclusion: we found that some of the variables given by this dataset are correlated with the chance of admit and explain a significant part of its variability; to validate this conclusion, we computed the feature importances using a trained ensemble model; the resulting model happened to outperform previously reported results over this dataset.

An important thing to mention is that the framing of the problem given by this dataset was extremely favourable to class purposes and it is very unlikely to find (in an easy way) this kind of dataset.

VI. Pending work

Use the selected model to compute a quantitative feature that captures the effect of feature importance, to evaluate how the attributes affect our target. This can be done using h2o built in functions and it will be used to validate the results found in the first stage of this project.

References

[1] Mohan S Acharya, Asfia Armaan, and Aneeta S Antony. A comparison of regression models for prediction of graduate admissions. In *2019 international conference on computational intelligence in data science (ICCIDS)*, pages 1–5. IEEE, 2019.